FaceLiveNet: End-to-End Networks Combining Face Verification with Interactive Facial Expression- Based Liveness Detection.

Zuheng Ming, Joseph Chazalon, Muhammad Muzzamil Luqman, Muriel Visani, Jean-Christophe Burie

FaceLiveNet: End-to-End Networks Combining Face Verification With Interactive Facial Expression-based Liveness Detection

Zuheng MING*, Joseph CHAZALON*†, Muhammad Muzzamil LUQMAN*,
Muriel VISANI* and Jean-Christophe BURIE*
*L3i, University of La Rochelle, France — †LRDE, EPITA, Paris, France
{zuheng.ming, joseph.chazalon, mluqma01, muriel.visani, jcburie}@univ.lr-fr

***Abstract*—The effectiveness of the state-of-the-art face verification/recognition algorithms and the convenience of face recognition greatly boost the face-related biometric authentication applications. However, existing face verification architectures seldom integrate any liveness detection or keep such stage isolated from face verification as if it was irrelevant. This may potentially result in the system being exposed to spoof attacks between the two stages. This work introduces FaceLiveNet, a holistic end-to-end deep networks which can perform face verification and liveness detection simultaneously. An interactive scheme for facial expression recognition is proposed to perform liveness detection, providing better generalization capacity and higher security level. The proposed framework is low-cost as it relies on commodity hardware instead of costly sensors, and lightweight with much fewer parameters comparing to the other popular deep networks such as VGG16 and FaceNet. Experimental results on the benchmarks LFW, YTF, CK+, OuluCASIA, SFEW, FER2013 demonstrate that the proposed FaceLiveNet can achieve state-of-art performance or better for both face verification and facial expression recognition. We also introduce a new protocol to evaluate the global performance for face authentication with the fusion of face verification and interactive facial expression-based liveness detection.**

I. Introduction

In the last few decades, face verification/recognition has been an active research topic. However, face verification/recognition is a challenge due to issues such as illumination conditions, variation of the pose, occlusion of the face, etc. [1]. Recently, the deep CNNs enabled to learn very effective high-level image features and significantly improved the state-of-the-art performance of the visual object classification or recognition problems [2]–[4]. Benefiting from the progress of visual object recognition, face recognition (as a sub-problem) has also made great breakthroughs such as the success of DeepFace [5], DeepIDs [6], VGG Face [7] and FaceNet [8]. Specifically, Facenet has firstly overpassed the human-level performance in terms of the accuracy on the benchmarks LFW [9] and YTF [10].

The great improvement of the state-of-the-art performance of face verification/recognition and the inherent convenience of face recognition have boosted the applications based on face biometric authentication, e.g. paying with the face, faceBioID as login information and so on. Accordingly, anti-spoof detection (also named liveness detection) is indispensable in practice for a face verification system to check for the actual presence of the user as opposed to a fake representation like a printed photo or a screen-shot from a video. However, the existing face verification architecture have either seldom enclosed any liveness detection, or kept this process as a separate stage irrelevant to the face verification task [11], [12]. Such face authentication framework which separates those two stages risks applying the liveness detection procedure to an individual different from the one who completed the face verification stage. Thus, a holistic end-to-end convolution neural networks namely FaceLiveNet is proposed to employ face verification and the liveness detection simultaneously (see Fig. 1). The liveness detection methods are of mainly two types: 1) static methods [11], [13], [14] and 2) dynamic methods, aka motion-based methods [12], [15], [16]. Static methods treat spoof detection as a binary classification problem based on the extracted features representing the texture differences (e.g. the specular reflection, image blurriness, image chromaticity and contrast distortion) between photo/video spoof and the real individual. However the generalization of the static methods is still limited between the different datasets (such as CASIA [17] , IDIAP [14] and MSU [11]). The inter-datasets evaluations reported in [11] show that in the case of the replay-attack test, the method LBP+SVM [13] gained only 7.8% in terms of the True Positive Rate (TPR) when False Accept Rate (FAR) equals to 0.1 if trained the model on CASIA and test on MSU, then as well the DoG+LBP+SVM [17] method gained only 14.2% of TPR when FAR equals to 0.1 if trained the model on IDIAP and test on MSU. Given that motion is a relative feature across video frames, the motion-based methods are expected to have better generalization ability than the texture based methods [15]. The motion-based methods

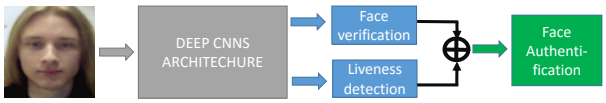


Fig. 1: The proposed end-to-end networks based on deep CNNs can simultaneously perform face verification and liveness detection for face authentication.

work by detecting the motion of the subject including the eye-blinking [12], the head pose [16] or the face motion [15] to perform liveness detection. However [18] shows that the simple motion such as eye-blinking and head rotation can be easily fooled with the crude photo-attack using images of the targeted person downloaded from social networks. Face motion detection based on the optical flow is also vulnerable to glasses, lighting conditions, and mustache which generates a distracting flow pattern [19]. Challenge-response authentication can be used to improve motion-based liveness detection. In this work, we propose to combine a standard facial expressions detection with a challenge-response mechanism to implement a liveness detection. Unlike the methods based on the 3D depth face or infrared face image which require expensive or uncommon sensors for anti-spoof detection, our method can be realized in the low-cost way using commodity hardware like smartphones. The challenge of this work is about performing efficient face verification and facial expression recognition in the same deep CNNs-based multi-task networks. In this work we propose to construct a deep networks based on the Inception-RsNet [20] structure with two main branches corresponding to the two tasks of face verification and facial expression recognition respectively. As well as face verification, the state-of-the-art performance of facial expression recognition has been significantly improved [21]–[23] owing to the deep CNNs. Unfortunately, the size of the existing datasets such as CK+ [24], OuluCASIA [25], SFEW [26] and FER2013 [27] are relative small comparing to the datasets used for face verification such as CASIA-Webface [28], MSCeleb-1M [29], etc. It is hard to train a deep CNNs from scratch for facial expression recognition with such small datasets. Inspired by the previous works [22], we leverage transfer learning to train our facial expression recognition branch from a pre-trained networks for face verification task. To the best of our knowledge, there is no dataset yet for the evaluation of system including simultaneously face verification and liveness detection based on interactive facial expression recognition. We therefore proposed a protocol to evaluate the global performance of FaceLiveNet for face authentication with the fusion of face verification and liveness detection, which can be used a baseline for the future work. Our main contributions are summarized as follows.

- We have proposed FaceLiveNet, a holistic end-to-end deep CNNs-based network which performs face verification and interactive facial-expression based liveness detection simultaneously for face authentication.
- We have introduced a protocol to evaluate the global performance of FaceLiveNet for face authentication.
- We have demonstrated that, for both face verification and facial expression recognition tasks, FaceLiveNet can achieve the stat-of-the-art or better performance on the datasets LFW, YTF, CK+, OuluCASIA, SFEW and FER2013.

In Section II we describe the architecture of the networks; in Section III we describe the training methods and Section IV presents the experimental results. Finally in Section V we draw the conclusions and present future directions of work.

TABLE I: The accuracies of facial expression recognition obtained by the FaceLiveNet with the different configurations of the Branch 2. The results are reported on the two datasets CK+ and OuluCASIA respectively.

	Block5 x1	Block5 x2	Block5 x2 + Block4
CK+	0.905	**0.991**	0.919
OuluCASIA	0.693	**0.875**	0.840

TABLE II: Number of the parameters included in the different deep CNNs architectures.

	FaceLiveNet	VGG16	AlexNet	FaceNet
Parameters	**1.31M**	138M	60M	7.5M

II. Architecture of FaceLiveNet

The FaceLiveNet is based on the Inception-RsNet structure. It is designed to have two main branches corresponding to the two tasks of face verification and facial expression recognition respectively as shown in (Fig. 2). The fusion of the results of face verification and facial expression recognition servers as the final result for face authentication. Specifically, the Branch 1 of FaceLiveNet extracts the embedded features for face verification and the Branch 2 calculates the probabilities of the facial expressions. The structure of the Branch 2 is almost same as the Branch 1 which can be easier for transfer learning. However, Table I shows that transferring more lower layers to Branch 2 have not consequently improved the performance for facial expression recognition. This may probably be explained by the lower layers inclining to learn the common features being less sensitive to the specific task. On the other hand, constructing the Branch 2 with too few layers, for example 1 block8 (4 layers) instead of 2 block8 (8 layers), weakens the representative capacity of the networks and results in the decline of the performance.

Although the FaceLiveNet based on the deep CNNs has about thirty layers, thanks to the several simplification techniques introduced by the Inception module, such as using the 1x1 convolution to reduce the dimension of the convolutions, and also factorizing the standard nxn convolution into 1xn and nx1 modules which reduce the grid-size of the networks while expanding the filter banks to keep the representation capability [20]. The total number of parameters of the network is only about 1.31 millions, which is much fewer than other popular deep CNNs such as VGG16 or FaceNet as shown in Table II. Having fewer parameters accelerates the training process. In practice, it takes about 12 hours to train the network on the CASIA-WebFace dataset with one Nvidia TitanX GPU for face verification, and then less than 1 hour to train the model for facial expression recognition on the CK+ or OuluCASIA datasets.

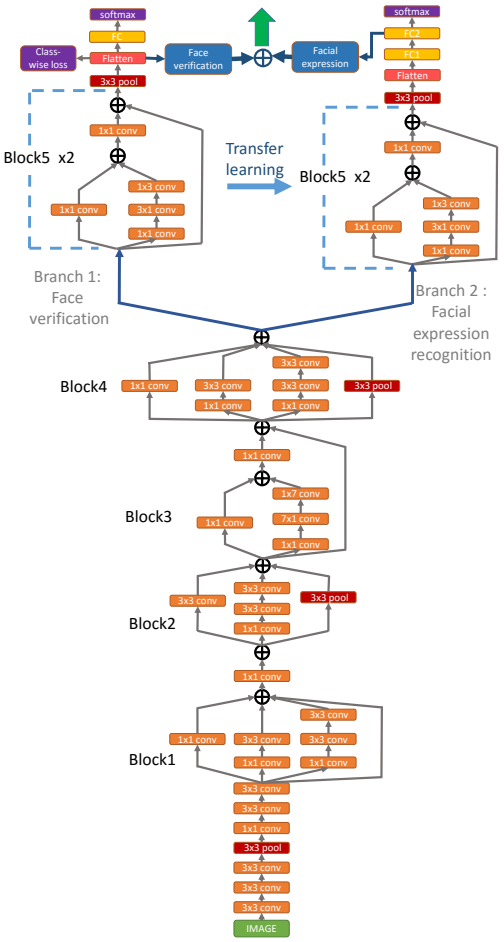


Fig. 2: The architecture of the proposed FaceLiveNet based on the Inception-RsNet structure which can perform face verification and facial expression recognition simultaneously with two main branches.

III. Approach and Training Protocol

The two branches are trained with different loss functions: Branch 1 is trained using class-wise triplet loss [30] joint with softmax loss for the task of face verification, and Branch 2 is trained using the softmax loss for the task of facial expression recognition. The class-wise triplet loss can be treated as a regularization term of the softmax loss during the training of Branch 1. The total loss L_1 of Branch 1 is given by:

$$L_1 = L_{s1} + \alpha L_c \tag{1}$$

where L_{s1} is the softmax loss, L_c is the class-wise triplet loss and the α is the weight of the class-wise triplet loss. The sofmax loss, i.e. the cross-entropy loss of the Branch 1 is given by:

$$L_{s1} = -\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y_i = j\} log \frac{e^{z_j}}{\sum_{l=1}^{k} e^{z_l}} \tag{2}$$

where m is the size of the mini-batch, k is the number of the classes, i.e. the number of the identities in the dataset, y_i is the label of the identity i. The class-wise triplet loss is given by.

$$L_c = \max(kD_{intra} + \beta - \theta D_\psi, 0) \tag{3}$$

Where, β and θ are the constants in terms of the distance margin and the weight of D_ψ respectively. D_{intra} is the sum of the intra-class distances of all the features $\mathbf{x_i}$ of the mini-batch to the center $\mathbf{c}_{y_i}$ of its class (i.e. the identities), D_{intra} is given by:

$$D_{intra} = \frac{1}{2}\sum_{i=1}^{m} \|\mathbf{x_i} - \mathbf{c}_{y_i}\|_2^2, \quad \mathbf{x_i}, \mathbf{c}_{y_i} \in \mathbb{R}^d \tag{4}$$

While D_ψ is the sum of the distances (including intra/inter-classes distances) of all the features $\mathbf{x_i}$ of the mini-batch to all the centers $\mathbf{c}_l$ of the different classes, then D_ψ is given by:

$$D_\psi = \frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{k} \|\mathbf{x}_i - \mathbf{c}_l\|_2^2, \quad \mathbf{x_i}, \mathbf{c}_l \in \mathbb{R}^d \tag{5}$$

Instead of averaging the features within class to calculate the center $\mathbf{c}_l$, it is updated dynamically with an initialization, which can avoid a perturbation of the values during the training:

$$\mathbf{c}_l^{t+1} = \mathbf{c}_l^t - \gamma \Delta \mathbf{c}_l^t \tag{6}$$

where t is the number of the iterations, and Δc_l^t is the variation of the centers during the updating, γ is the learning rate for updating. The variation of the center $\Delta \mathbf{c}_l$ is given by:

$$\Delta c_l^t = \frac{\sum_{i=1}^{m} 1\{y_i = l\} \cdot (\mathbf{c_l}^t - \mathbf{x_i}^{t+1})}{\sum_{i=1}^{m} 1\{y_i = l\}} \tag{7}$$

The loss for Branch 2 is simply based on the softmax loss L_{s2}. The equation of L_{s2} is same as L_{s1}. The only difference is that the k in L_{s2} is the number categories of the facial expressions, e.g. 6 or 7 expressions rather than the number of identities.

The details of the training protocol is summarized in Algorithm. 1.

IV. Experimental Results and Analysis

In this section, the proposed FaceLiveNet is evaluated firstly for face verification and facial expression recognition tasks respectively. Then we propose a dataset to evaluate the global performance of FaceLiveNet for face authentication with the fusion of face verification and facial expression recognition.

A. Evaluation for face verification

Table III shows the evaluation results of FaceLiveNet on the datasets LFW and YTF. The model has been trained on CASIA-WebFace and MSCeleb-1M respectively. In both of the training and evaluation phase, the face images have been detected by the MTCNN [31]. Specially, some data augmentation processing such as the random crop, random flip and data filtering have been applied during the training phase. The SGD and the mini-batches of 90 samples are applied

Algorithm 1 Training Protocol of FaceLiveNet

Stage 1: Training Branch1 and the main stem of networks for face verification task

Input: Training samples $\{\mathbf{I_{1,i}}\}$ for face verification

Output: The Branch1 and main stem networks parameters $\{\mathbf{w_1}\}$

1: Initialize the parameters by Xavier Initialization
2: **for** $i = 1 : T_1$ **do**
3: Forward propagation: the total loss $L_1 = L_{s1} + \alpha L_c$
4: Update the centers: $\mathbf{c}_l^{t+1} = \mathbf{c}_l^t - \gamma \Delta \mathbf{c}_l^t$
5: Back propagation: update the parameters of the networks $\mathbf{w_1}^{t+1} = \mathbf{w_1}^t - \lambda^t (\frac{\partial L_{s1}}{\partial \mathbf{x_{1,i}}} \cdot \frac{\partial \mathbf{x_{1,i}}}{\partial \mathbf{w_1^t}} + \frac{\partial L_c}{\partial \mathbf{x_{1,i}}} \cdot \frac{\partial \mathbf{x_{1,i}}}{\partial \mathbf{w_1^t}})$
6: **end for**

Stage2: Training Branch 2 for the face expression recognition task

Input: Training samples $\{\mathbf{I_{2,i}}\}$ for the face expression recognition

Output: The Branch 2 networks parameters $\{\mathbf{w_2}\}$

1: Freeze the Branch1 and the main stem of the networks
2: Initialize the Branch2 by the pretrained Branch1
3: **for** $i = 1 : T_2$ **do**
4: Forward propagation: the total loss L_{s2}
5: Back propagation: update the parameters of the networks $\mathbf{w_2}^{t+1} = \mathbf{w_2}^t - \lambda^t (\frac{\partial L_{s2}}{\partial \mathbf{x_{2,i}}} \cdot \frac{\partial \mathbf{x_{2,i}}}{\partial \mathbf{w_2^t}})$
6: **end for**

Method	Images	Nets	LFW	YTF.
Fisher Faces [33]	-	-	93.10	83.8
DeepFace [5]	4M	3	97.35	91.4
DeepID-2,3 [6]	-	200	99.47	93.2
FaceNet [8]	200M	1	99.63	95.1
VGGFace [34]	2.6M	1	98.95	91.6
Centerloss [35]	0.7M	1	99.28	94.9
FaceLiveNet (CASIA)	0.46M	1	**98.91**	**94.88**
FaceLiveNet (MSCeleb)	1.1M	1	**99.42**	**95.00**

TABLE III: Evaluation results of FaceLiveNet for face verification on the LFW and YTF datasets.

for training the deep CNNs in this work. The momentum coefficient is set to 0.99. The learning rate is started from 0.1, and divided by 10 at the 60K, 80K iterations respectively. The model is regularized by using the dropout with the probability of 0.5 and the weight decay of 5e-5. The weights of the filters in the CNNs are initialized by Xavier [32]. Biases are initialized to zero. The weight of the class-wise triplet loss α is set to 1e-4, the margin β is set to 10, and the weight of the inter-class distance θ in the class-wise triplet loss function is set to 0.5. This evaluation shows that the FaceLiveNet can achieve the state of art on the benchmarks LFW (99.42%) and YTF (95.00%) for face verification task, while a larger dataset helps to improve the performance.

B. Evaluation for facial expression recognition

Facial expression recognition is evaluated on the four widely used datasets: CK+, OuluCASIA, SFEW and FER2013 respectively. The CK+ and OuluCASIA are the datasets for the posed expression while SFEW and FER2013 are the datasets collected from website in an unconstrained condition. The distribution of the expression images of the four datasets is shown in Table IV. As in [36], the last three frames and the

TABLE IV: The distribution of the expression images of the different datasets: Neutral (Ne), Anger (An), Disgust (Di), Fear (Fe), Happy (Ha), Sad (Sa), Surprise (Su), Contempt (Co).

	Ne	An	Di	Fe	Ha	Sa	Su	Co
CK+	327	135	177	75	147	84	249	54
OuluCASIA	-	240	240	240	240	240	240	-
SFEW	228	225	75	124	256	234	150	-
FER2013	4965	3995	436	4097	7215	4830	3171	-

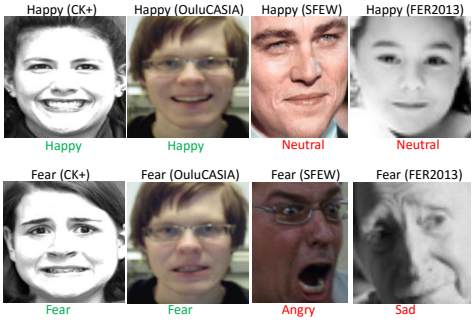


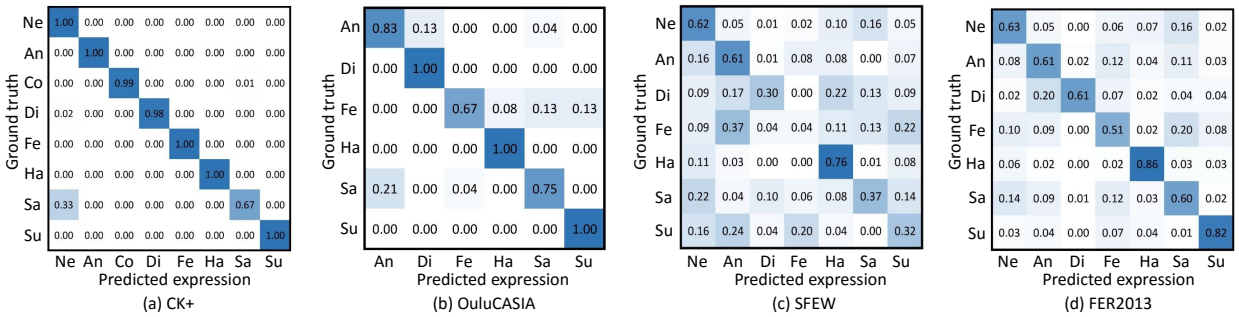
Fig. 3: The samples of facial expression recognition results on the different datasets. The ground truth labels are on the top of the images and the predicted labels are on the bottom.

first frame of each video of CK+ and OuluCASIA are selected for the training and evaluation. The 10-folds cross-validation with the subject independent split is applied for the evaluation on CK+ and OuluCASIA. Table V demonstrates the evaluation results on the four datasets and Fig. 4 shows the corresponding confusion matrix of facial expression recognition.

The evaluations on the four different datasets show that the proposed FaceLiveNet can also achieve the state-of-the-art for facial expression recognition task. However the recognition results on CK+ and OuluCASIA are much better than the ones on SFEW and FER2013. Since CK+ and OuluCASIA are collected in a constrained condition with the distinct expression and clean data. While SFEW and FER2013 are collected from the website or the film clip, the expressions are more delicate and the representation of some expressions varies greatly depends on the habits of the individuals (see Fig. 3). Overall, according to the results, the 'Happy' and 'Surprise' are the expressions most universal and always with the highest recognition rate, while the 'Fear' and 'Sad' are vulnerable to be mispredicted. Thus the 'Happy' and 'Surprise' are used as the required expressions by the system for liveness detection based on challenge-response mechanism.

C. Evaluation for face authentication

In this section, we propose a protocol to evaluate the proposed FaceLiveNet. Unlike the general facial expression detection problem focusing on the spontaneous expression in the real-life, in the challenge-response based liveness detection the facial expressions are required to be presented distinctly which are close to the ones in CK+ or OuluCASIA. Thus we use the images from the two datasets to construct the image



(a) CK+ (b) OuluCASIA (c) SFEW (d) FER2013

Fig. 4: The confusion matrix of facial expression recognition results for (a) CK+, (b) OuluCASIA, (c) SFEW, (d) FER2013. The vertical axis is the ground truth label and the horizontal axis is the predicted expression. The darker color the higher value.

Method	Acc(%)
LBPSVM [37]	95.1
Inception [23]	93.2
DTAGN [21]	97.3
PPDN [22]	97.3
FaceLiveNet	**99.1**
AUDN [38]	92.1
FaceLiveNet	**98.0**

(a) CK+

Method	Acc(%)
HOG3D [39]	70.63
AdaLBP [25]	73.54
DTAGN [21]	81.46
PPDN [22]	84.59
FaceLiveNet	**87.50**

(b) OuluCASIA

Method	Acc(%)
AUDN [38]	31.73
MappedLBP [40]	41.92
Inception [23]	47.70
FaceLiveNet	**49.50**
TL [41]	48.50
MDNN [42]	52.29
FaceLiveNet	**53.20**

(c) SFEW

Method	Acc(%)
RBM [39]	71.162
Unsupervised [27]	69.267
Maxim [27]	68.821
Radu [43]	67.484
Baseline [27]	65.500
FaceLiveNet	**68.600**

(d) FER2013

TABLE V: The evaluation of FaceLiveNet for facial expression recognition on different datasets. Specially, in (a) the upper block are the results of six expressions in CK+ and the lower block are the results for the eight expressions; in (c) the upper block are the results by using SFEW training the model, and the lower block are the results by using FER2013 as the additional data for training the model.

	ID-True	ID-False
Ex-True	38	56
Ex-False	56	56

(a) CK+

	ID-True	ID-False
Ex-True	96	96
Ex-False	96	96

(b) OuluCASIA+

TABLE VI: The positive pairs and negative pairs in the evaluation dataset for face authentication extracted form CK+ and OuluCASIA.

	Acc_{verif}	Acc_{expre}	Acc_{live}	Acc_g
(CK+)	0.990	0.981	1.000	0.990
(OuluCASIA)	0.932	0.935	0.990	0.922

TABLE VII: The evaluation results of face authentication of FaceLiveNet on the dataset CK+ and OuluCASIA. Acc_{verif} is the accuracy of face verification, Acc_{expre} is the accuracy of facial expression recognition, Acc_{live} is the accuracy of liveness detection and Acc_g is the accuracy of the face authentication.

pairs to simulate face authentication scenario: the first image is used as the answer of the individual and the second one is the neutral image used as reference image. The two images are compared firstly for face verification and then the detected facial expression of the first image is compared with the system required expression for liveness detection. The dataset consists of the positive pairs and the negative pairs as shown in Table VI. Note that the data used for the evaluation are not used for the training of the model. Totally, 206 image pairs are extracted from CK+ while 384 images pairs are extracted from OuluCASIA. Finally, the accuracy of FaceLiveNet for face authentication Acc_g is given by:

$$Acc_g = \frac{\sum_{i=1}^{M} \{V_i \cap E_i\}}{M} \tag{8}$$

where M is number of the image pairs, $V_i \in (True, False)$ is the result of face verification for the ith pair, $E_i \in (True, False)$ is the result of liveness detection based on the verification of the given facial expression 'Happy' or 'Surprise', $\{\cdot\}$ is the indicator function. Table VII illustrate the accuracy of FaceLiveNet for face authentication. From the Table VII we can see that the accuracy of the liveness detection is higher than facial expression recognition rate since liveness detection only detects the given expressions 'Happy' or 'Surprise' rather than the recognition of the six or eight expressions. However, the global accuracy of face authentication can be also worse than both face verification and liveness detection, which is caused by the additive effect of the failures of face verification and liveness detection.

V. Conclusion

In this work, we have proposed a holistic end-to-end networks based on the deep CNNs which can employ simulta-

neously face verification and facial expression recognition for face authentication. Experimental results demonstrate that for both face verification and facial expression recognition tasks FaceLiveNet can achieve the state of art. Besides, a protocol is proposed firstly for the evaluation of face authentication by FaceLiveNet with the fusion of face verification and liveness detection. Thanks to the Inception-RsNet structure, the proposed FaceLiveNet is light with fewer parameters in compare with the other conventional deep networks such as VGG16, FaceNet etc. and also low-cost as it can implement on commodity hardware instead of costly sensors. In future work, we will develop a larger dataset to evaluate the performance of our proposed FaceLiveNet.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the project of MOBIDEM.

REFERENCES

[1] W. Zhao, R. Chellappa, and et al., "Face recognition:a literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[3] C. Szegedy, W. Liu, and et al., "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[4] K. He, X. Zhang, and et al., "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[5] Y. Taigman, M. Yang, and et al., "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014, pp. 1701–1708.

[6] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *CVPR*, 2015, pp. 2892–2900.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[10] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR*, 2011, pp. 529–534.

[11] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[12] G. Pan, L. Sun, and et al., "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *ICCV*, 2007, pp. 1–8.

[13] T. de Freitas Pereira and A. et al., "Lbp-top based countermeasure against face spoofing attacks," in *ACCV*, 2012, pp. 121–132.

[14] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2012, pp. 1–7.

[15] S. Bharadwaj, T. I. Dhamecha, and et al., "Computationally efficient face spoofing detection with motion magnification," in *CVPR Workshops*, 2013, pp. 105–110.

[16] R. W. Frischholz and A. Werner, "Avoiding replay-attacks in a face recognition system using head-pose estimation," in *International Workshop on Analysis and Modeling of Faces and Gestures*, 2003, pp. 234–235.

[17] Z. Zhang, J. Yan, and et al., "A face antispoofing database with diverse attacks," in *International Conference on Biometrics*, 2012, pp. 26–31.

[18] Z. Boulkenafet, J. Komulainen, and et al., "Oulu-npu: A mobile face presentation attack database with real-world variations," in *International Conference on Automatic Face & Gesture Recognition*, 2017, pp. 612–618.

[19] K. Kollreider, H. Fronthaler, and J. Bigun, "Verifying liveness by multiple experts in face biometrics," in *CVPR Workshops*, 2008, pp. 1–6.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[21] H. Jung, S. Lee, and et al., "Joint fine-tuning in deep neural networks for facial expression recognition," in *ICCV*, 2015, pp. 2983–2991.

[22] X. Zhao, X. Liang, and et al., "Peak-piloted deep network for facial expression recognition," in *ECCV*, 2016, pp. 425–442.

[23] A. Mollahosseini, D. Chan, and et al., "Going deeper in facial expression recognition using deep neural networks," in *Winter Conference on Applications of Computer Vision*, 2016, pp. 1–10.

[24] P. Lucey, J. F. Cohn, and et al., "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *CVPR Workshops*, 2010, pp. 94–101.

[25] G. Zhao, X. Huang, and et al., "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29.

[26] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *International Conference on Multimodal Interaction*, 2015, pp. 423–426.

[27] I. J. Goodfellow, D. Erhan, and et al., "Challenges in representation learning: A report on three machine learning contests," in *NIPS*, 2013, pp. 117–124.

[28] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[29] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016, pp. 87–102.

[30] Z. Ming, J. Chazalon, and et al., "Simple triplet loss based on intra/inter-class metric learning for face verification," in *CVPR Workshops*, 2017, pp. 1656–1664.

[31] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[32] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[33] K. Simonyan and O. M. e. a. Parkhi, "Fisher vector faces in the wild." in *BMVC*, vol. 2, no. 3, 2013, p. 4.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016, pp. 499–515.

[36] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *CVPR*, 2014, pp. 1749–1756.

[37] X. Feng, M. Pietikäinen, and A. Hadid, "Facial expression recognition based on local binary patterns," *Pattern Recognition and Image Analysis*, vol. 17, no. 4, pp. 592–598, 2007.

[38] M. Liu, S. Li, and et al., "Au-aware deep networks for facial expression recognition," in *International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–6.

[39] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, 2008, pp. 275–1.

[40] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *International Conference on Multimodal Interaction*, 2015, pp. 503–510.

[41] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *International Conference on Multimodal Interaction*, 2015, pp. 443–449.

[42] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *International Conference on Multimodal Interaction*, 2015, pp. 435–442.

[43] R. T. Ionescu, M. Popescu, and C. Grozea, "Local learning to improve bag of visual words model for facial expression recognition," in *Workshop on challenges in representation learning, ICML*, 2013.